H14-223

USE OF NEURAL NET MODEL TO FORECAST CARDIOVASCULAR AND RESPIRATORY DISEASES FROM METEOROLOGICAL AND POLLUTION DATA.

Armando Pelliccioni[1] and Rossana Cotroneo[2]

[1]Ispesl-Dipia, Via Fontana Candida 1, 00040 Monteporzio Catone, Italy
[2]Istat, Viale Liegi 13, 00198 Roma, Italy

Abstract: One of the major environmental problems is air pollution that has important effects the quality of living and health conditions of the population especially in urban areas. According tothe World Health Organization, in Europe air pollution is the eighth cause of death and is the main environmental risk factor (World Health Organization). Moreover, the issue of air quality is now a major concern for many governments worldwide.
Air pollution depends by number and diversity of emission sources and by concomitant risk factors, taking into account geographical and meteorological conditions do not permit an easy circulation of air and a large part of the population moves frequently between distant places of a city.
Such situations are very important especially for the Mediterranean coastal countries (like Italy) and for urban places that are affected by the microclimatic conditions that could lead to atmospheric stagnation events.
The aim of this work is to analyse the effects of the particulate matter (PM_{10}) and ozone (O_3) on cardio-respiratory diseases by neural network (NN).
In our analysis we introduced some important variables to NN training, such as the skewness and mean daily of the distribution of PM_{10} and Ozone, as well as the temperature and relative humidity.
The results coming from NN model to investigate in short-term the relationship between air pollution, mortality and morbility, shows that it is possible to predict the pollution's impacts in terms of number of hospital admissions related to cardio-respiratory system.

Key words: *air pollution, ozone, particulate matter, neural network,times series, forecasting, cardio-pulmonary disease,hospital admissions; meteorology*

INTRODUCTION

The aim of this paper is to provide a contribution in the field of air pollution especially due to particulate matter and ozone concentrations. PM_{10} and O_3 that tends aretends to have stronger effects in the summer during periods of higher concentration, significantly related to the degradation of air quality in urban agglomerations, generating adverse health effectswith increased use of hospital services. Indeed, air pollution not only affects health care systems, as we cited above, but even local economy in terms of costs of medication, absences from work, and child and old people care expenses. In particular, the purpose of this study is to determine a way of assessing exposure to outdoor pollution.

In the short term (the data used covered a period of two years: 2005 and 2006) their impact assessment on population (especially on children, the patients, pregnant women and elderly and on people who suffer from chronic cardio-respiratory diseases) in terms of morbidity and mortality by cardio-pulmonary diseases (Murray A. and M.D. Mittleman, 2007) in the urban area of Rome,whose population and industrialization is continuously increasing.

Our aim is to analyze the annual trend of morbidity and mortality in hospital for cardio-respiratory diseases due to PM_{10} and O_3 by using neural net (NN) techniques (C.M. Bishop, 1995), which are important tools quickly and efficiently to forecast due to their universal approximation property (i.e. the capability of approximating a non linear function as precisely as needed, by increasing the number of parameters) and their fast training (if sequential training based on back-propagation in adopted), without needing any distribution or relationship between variables and a priori information about the physical causes of air pollution.The NN capacity to learn non-linear functions is an important issue in our problem. Moreover, neural network does not require a priori assumptions about the input variable distribution or absence of correlations between such variables. Consequently, NN can be used in evaluating cardio-respiratory diseases due to environmental systems (R. Rojas, 1996) and can capture without many of the usual limiting assumptions of other traditional advanced statistical methods, which make not particularly appropriate.

The NN was trained to predict up to five days in advance the number of mortality in hospital for cardio-respiratory diseases per day based on the input of meteorological and air pollution data.Thus it will be possible to increase the knowledge of levels of concentration of pollution in urban areas and inform the public about environmental risks on human health.By applying NN in the field of environmental epidemiology (World Health Organisation, 2004) we want to:

- examine the most injurious pollutants to human health and their mechanisms of dispersion
- identify factors and mechanisms of epidemic diffusion
- model the cardiovascular and respiratory diseases by means the use of air pollution data coming from urban monitoring station
- formulate a methodological suggestion to apply Neural Net to forecast five days in advance human health impact due to the different pollutants
- provide a reliable model for the prediction of the daily hospital admissions based on air quality and meteorological data, undoubtedly useful for regulatory purposes.

MATERIAL AND METHODS

In Rome, to measure the effects on population health due to exposure to physical, chemical and biological agents outside the human body, we analysed data of the hospital discharge records (HDR) for the calendar years 2005 and 2006. In public

health, the HDRs are the tool of information gathering for every patient discharged from public and private hospitals throughout the territory. They provide relevant information about the diagnosis. This information is collected in a data warehouse (EPICS) by Agency of Public Health (ASP) of Lazio.

Instead, to measure the air pollution levels we analysed data related to primary pollutants and meteorological variables cover the 2005 and 2006 period and coming from monitoring stations of the ARPAL (Environmental Protection Agency of Lazio Region).

Our environmental dataset examines about 90,000 hourly pattern data, coming from 5 monitoring stations recorded at different urban sites, and is composed by pollutants variables and conventional meteorological variables:

- observed air pollutants variables:
  - Carbon monoxide (mgm^{-3}) – CO-
  - Nitrogen Oxide (μgm^{-3}) –NO-
  - Nitrogen Dioxide (μgm^{-3}) -NO2-
  - Mono-Nitrogen Oxides (μgm^{-3}) -NOx-
  - Ozone (μgm^{-3}) - O_3 -
  - PM_{10} (μgm^{-3}) - PM_{10}-

- meteorological variables:
  - Temperature (C°) – T-
  - Global Solar Radiation (Wm^{-2}) – GSR-
  - Relative Humidity (%) – RH-
  - Pressure (mbar) – P-
  - Rain (mm).

Moreover, we observe that our environmental time series contain information at different scales, some of which are periodic, such as day and hour that represent seasonality and diurnal variation. These variables were treated in term of cosine and sine. In this way, it is possible to take into account the effects of seasonality and diurnal variation.The input data used for prediction cardio-pulmonary morbidity and mortality was the average of the recordings of all the stations, after we determined that there were no significant differences among them. Before conducting any analysis, previously all data were standardised (pollutants and meteorological conditions).

To analyse the relationship between air pollution and cardio-respiratory diseases, we examined the following statistical characteristics of the distribution of hourly data specified by its moments: mean, standard deviation, skewness (Sk) and kurtosis, that refers to the peak of a frequency curve, maximum, minimum and variation filed values calculated on the chemical and meteorological parameters.

Skewness is a measure of the symmetry of the shape of the distribution of our variables. If a distribution is symmetric, the skewness will be zero. If there is a long tail in the positive direction, skewness will be positive, whereas if there is a long tail in the negative direction, skewness will be negative.In our paper, SK allowed to underline in a significant manner our aim. In particular, the analysis of stratified data by day allow you to see what form it takes the distribution of different variables at each calendar day and verify the existence of particular trends in the time series associated with these layers of time. In particular, it was possible to observe a not linear relation between averages of CO's daily distributions and their skewness, of the CO's daily distributions and averages of O_3's daily distributions and their skewness. These relations allow explaining in a good manner the latent factors that are not manifest variables. For this reason, we included these indices as input for NN, because the greatest advantage of a neural network is its ability to model a complex non-linear relationship between independent and dependent variables (M.W. Gardner and S.R. Dorling, 1999),(Gardner and S.R. Dorling, 2000),(S.A.Abdul-Wahab and S.M. Al-Alawi 2002).

Moreover, we calculated the daily distributions (PDF) of each pollutant (760 PDFs for each variable), for confirming the hypothesis of a concentration trends related to changes in the day. We synthesized these PDFs through their mean, standard deviation and skewness that are able to capture environmental performance. If a daily distribution presents a negative asymmetry it means that its values will be below of alert values for the exposure of population.

METHODOLOGIES

Some preliminary retrospectives analysis were carried out on hourly environmental time series (F. Battaglia, 2007) (pollutants and meteorological variables) and on daily time series related to admissions for cardio-respiratory disease, identifying and examining graphically their trends. Such time series constitute a summary of the "history" of the environmental phenomenon investigated. Thus we proceeded first to examine the fundamental characteristics of the series.
Then we carried on detecting the most relevant periodic components for the explanation of the variability of the series by a spectral analysis. The spectral analysis consents to highlight what are the frequencies (and hence the periodicity) more important. The periodogram, in fact, measures the intensity of k-frequency within the range of values and hence the importance that assume each period p_k.

Based on the results coming from the preliminary analysis of these series, was designed and developed a simulation model of interpretation implemented by neural network. NN was developed to provide five days in-advance forecasts of the cardio-pulmonary disease.

NN are computer-based algorithms inspired by the structure and behaviour of real neurons. Like the brain, they can recognize patterns, reorganize data and, most appealing, and learn from experience. The NN consists of a set of processing units that simulate neurons and are interconnected via a set of weights in a way that allows signals to travel in parallel as well as serially. The greatest advantage of a neural network is its ability to model a complex non-linear relationship (M.W. Gardner and S.R. Dorling, 1999) (Gardner and S.R. Dorling, 2000), (S.A.Abdul-Wahab and S.M. Al-Alawi 2002) by traditional approaches, such as those in the environmental systems, without a priori assumptions on its nature (S. BuHamra et al, 2003), by means of an accurate choice of the variables of the system and of the meaningful patterns, and data distribution.

For transfer function, the most suitable architectures are considered to be the Multi Layer Perceptron (MLP) (H. Abdi, 1994); (J.B. MacQueen, 1967) (L. Fausett, 1994); (C.M. Bishop, 1995); (B.D. Ripley, 1996) that are biologically inspired neural models consisting of a complex network of interconnections between basic computational units, called neurons and with an error-back-propagation supervised learning rule (Rojas, 1996). Of several possible ways to train an NN, one of the most successful supervised training methods is the back-propagation algorithm. The basic concept is to use the derivative of an error function in order to find the direction that minimizes the error of the network and updating the weights accordingly. The algorithm attempts to minimize the mean error over the entire training set.

This net architecture is able to reproduce non linear models, without any a priori assumptions, by means of an accurate choice of the variables of the system and of the meaningful patterns. A learning algorithm is an adaptive method by which a network of computing units self-organises to reproduce the desired model. This is done with learning algorithms that present some examples of the desired input-output mapping to the network. A correction step (the error-backpropagation rule) is performed iteratively until the network learns to produce the desired response.

As architecture we used a 3-layer perceptron model with a single hidden layer, 10 hidden neurons and with sigmoid activation function (see equation 1) that approximates nonlinearities (see table1).

$$F(P) = \frac{1}{1 + e^{-(p-s)}} \qquad (1)$$

where P is the activation potential and S is the activation threshold.

The first inputlayer contains the input variables of thenet, pollutants and meteorological variables. The second layer consists of the neurons of the hidden layer. The third layer is the output layer, which consists of the target of the forecasting model.
The number of neurons of the hidden layer is one of the parameters to be chosen in the NN model architecture, the well known multi-layer perceptron.

The choice of 10 hidden neurons is based on two considerations: maximizing the hidden neurons to increment the NN parameters and simultaneously minimizing this number in relation with the main situations linked to the input patterns. Moreover, we utilise different methods to optimize the weight values of hidden layers, in the manner that the errors of the network's output could be minimized.

Table1. Neural Networks architecture

NEURAL NETWORK MODEL	MLP 6-10-1
HIDDEN NEURONS	8-10-12
ALGORITHM	CG-BFGS-GD
EPOCH	3000
ERROR FUNCTION	SUM OF SQUARE
HIDDEN ACTIVATION FUNCTION	LOGISTIC-TANH
OUTPUT ACTIVATION FUNCTION	IDENTITY
NETWORK RANDOMIZED	NORMAL

The nature of the functional relationship between inputs and outputs is learnt during a supervised training process directly from the data. Neural Networks can be trained to accurately generalize when presented with new, unseen data. Often, especially in the atmospheric sciences, successfully modeling the average behavior of a system is not the main goal. It is important sometimes that the model can also interpret infrequent outliers, which are often of great importance, as for the health related with exposure.

In this work, we underlined also the question of the best variables of input to be used and the optimum selection of patterns for having results that represent the health response to the different pollutants and to meteorological factors.
At the end, to compare NN result we analysed the results coming from auto-regressive model that assess the linear nature of the relationship between the dependent and independent datasets. The main disadvantages consist in transformation of non-linear relations are into linearity and the existence of multicollinearity. In fact, the best auto-regressive model could become a particular case (the linear limit) of NN.

RESULTS AND DISCUSSION

The effects of PM_{10} and ozone linked to short-term exposure on cardio-respiratory diseases have been documented by the joint study of time series that examines changes of health outcomes related to changes in concentration levels of pollutants. To this end, the distributions daily of O_3 and PM_{10} and hospital admissions were examined in terms of morbidity and mortality. Thus, it was possible to verify the existence of dynamic relationships between time series of pollutants and the time series of hospitalizations for cardiopulmonary disease.

In terms of morbidity, it was observed a possible causal link between the respiratory and cardiovascular diseases.

For ozone, the results show a meaningful increase in daily hospital admissions for cardio-respiratory system due to high concentration levels of this pollutant.

In terms of mortality for respiratory disease, it was observed that at high of concentration levels of PM_{10} increase the number of deaths. The relationship, however, is less evident than for cardiovascular disease, as there is found a lower proportion of deaths. Results for ozone show significant increases in deaths related to exposure to high levels of concentration.

Besides carrying out the analysis of trends of pollutant concentration levels and hospital admissions, we proceeded with the description of the main results obtained with the neural network model. The daily PM_{10} and O_3 data and the aggregated data for cardio-respiratory disease (ICD9CM 390-519) that relate to hospital were taken into consideration.

From our results, we can observe that the neural network is able to reproduce a good approximation the causal link between the concentration levels of PM_{10} and of O_3 and admissions (World Health Organisation, 2006).

Epidemiological data are used to investigate the relationship between PM_{10} and O_3 and morbidity and mortality, suggests that most of the differences in mortality can be attributed to a worsening of pre-existing conditions rather than the onset, due to pollution air pollution, of new diseases.As suggested by above considerations, the performances of the NN or the ability to reproduce the target variables (in our case the hospitalisation for cardiovascular and respiratory diseases), are strictly linked to the variables and patterns selections. As usually, we use 65% of random patterns during the training and the remaining 35% as test, never seen by NN during the learning phase.Our aim is that we attempt to reproduce the average five days in advance of hospitalisation for the two considered pathologies given as input data best subset of the following variables: general information (Julian day and week Day), air quality data (pollutants) and meteorological data (see Table.2).

Table 2: NN input variable

NN Input Variables		
Julian Day	Sin (Jd)	Cos (JD)
Week Day	Sen (Week)	Cos (Week)
Pollutants	PM_{10}	
	NO2	
	CO	Sk (CO)
	O_3	Sk (O_3)
Meteorological	T	HR

In order to facilitate the comparison of NN modelling, we show results related to the cardio-respiratory hospitalisations separately.For cardiac disease, the NN predicted well the hospitalisations during all the days of period (the calendar years 2005-2006). The determination coefficient is very high (R^2=0.93) and we underestimate the extremes values of admissions both in term of the lower than upper limits.For respiratory disease, the determination coefficient is a little worse respect to previous simulation (R^2=0.92), even if levels are very significant.At the end, we compared the conventional regression model with neural networks to forecast cardio-respiratory diseases under different meteorological and pollution variable. This model performs less satisfactory than NN perform, especially for cardiac admissions, with an R^2 of 31% and 76% for respiratory admission.

CONCLUSION

This study focused on NN methodology to investigate the short-term impact of air pollution on public health in urban areas, as functions of some chemical variables and of local meteorological parameters.In particular, we have developed a methodology to build neural networks that accurately predict the cardio-respiratory disease by using information from environmental air pollutants and atmospheric data. In fact, the network captured the complex correlation between the observed variation in air pollution, weather conditions and cardio-respiratory disease.

A good model has been developed in terms of a high correlation coefficient between actual and predicted values. Moreover, the robustness of the models is guaranteed, since forecasting is accurate throughout the entire predictionperiod.Results obtained confirm the utility of similar neural architectures for predicting the health impact of air quality pollutant concentrations.

Our data show that in Rome the concentration of PM_{10} and ozone have exceeded the danger levels suggested by the European Guidelines (DIR2008/50/CE).One of more interesting aim of for the health and environmental question are the connections between the air quality data and the relative effects on the human health, i.e. the hospitalisation. In general, while is well know that air quality impacts on the health, the quantification of this the effect is hard to simulate. This happen because the

relation between the main variables (such as meteorological and pollutants ones) and health effects cannot be determined directly by deterministic models or by simplified statistical models. In fact, we can suppose that this relation could be non linear type and that the role of the important variables is hidden by the complexity of the environmental phenomenon investigated.In this context, we use one of the most advanced non linear models, such as the Neural Network, to attempt the relations between the environment and meteorological data and health effects on the populations. To this regards, we investigate the cardio-respiratory hospitalisations for total population of the Rome city, during the 2005 and 2006.

To that regards, we found that the conventional statistical descriptors, such as the daily average of pollutants, often cannot be link to the exposure levels, if they are taken standalone. By the pre-processing analysis, we demonstrated that the skewness coefficients for the pollutants can give a more accurate connection with the real human exposure.In our work, we applied an intelligent models constituted by a Neural network model to connect the environmental data with the hospitalisations for the two considered diseases.The results obtained by NN model are very encouraging and suggest a way to modelling this complex relation. In fact, we obtained a very meaningful correlations (higher than 0.90) for both simulations. While the cardiac pathology is better reproduced by NN, the respiratory ones have needed more analysis in deep. However, both simulations exhibit that, to optimize the training phase, the choice of the input variables and the choice of patterns are the main factors to be considered for successful of intelligent methodology.

In our study, it is evident that the performances obtained is link to the right choice of input factors and that the NN performance is good only if some heavy pre-processing evaluation are given on input data.These first results showed the importance of the environmental-epidemiological problem and the major areas in order to forecast with some accuracy the short-term effects on human health of the environmental component. In particular, the forecasting with up to five days in advance would allow taking more efficient countermeasures to safeguard citizens' health.By using neural networks, it was possible to determine numerically the association statistically meaningful between the cases of death or hospitalization and pollutants (PM_{10} and/or O_3).

The results coming from this work have used neural networks to investigate short-term the relationship between air pollution, mortality and hospital admissions, identifying and evaluating the sources of pollution in order to define and adopt effective mitigation measures for air quality improvement of Rome and the promotion of strategies for the prevention and treatment of cardio-respiratory diseases.

As last consideration, we underline that the results obtained seem encouraging to simulate the effects of air quality on the health through NN model, but at the same time indicate that further study are necessary in future as to extend the NN's prediction to more years and to consider also the spatial distribution of pollutant in relation with the local hospitalisation.

REFERENCES

Abdul-Wahab, S.A.and S.M. Al-Alawi, 2002: Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environmental Modelling & Software* 17, 219-228pp.

Abdi, H.,1994 : Les re´seaux de neurones. Presse Universitaire de Grenoble

Battaglia F.,2007: Metodi di Previsione Statistica. Springer Verlag

Bishop, C.M., 1995: Neural Networks for Pattern Recognition. Clarendon Press, Oxford

BuHamra, S., Smaoui, N., Gabr, M. (2003). The Box-Jenkins analysis and neural networks: prediction and time series modelling. *Applied Mathematical Modelling* 27 (10), 805-815pp.

Fausett, L. (1994). Fundamentals of Neural Networks. In: *Architectures, Algorithms and Applications*. Prentice Hall, Englewood Cliffs, NJ 07632

Gardner, M.W. and Dorling, S.R. (2000). Statistical surface ozone models: an improved methodology to account for non-linear behaviour. *Atmospheric Environment* 34, 21-34pp.

Gardner, M.W.and Dorling, S.R. (1999). Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. *Atmospheric Environment* 33, 709-719pp.

MacQueen J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, University of California Press, 1:281-297pp.

Murray, A. and M.D. Mittleman (2007). Air Pollution, Exercise, and Cardiovascular Risk. *The New England Journal of Medicine*,2007

Ripley, B.D.,1996: Pattern Recognition and Neural Networks. CambridgeUniversity Press

Rojas R., 1996: Neural Networks: a systematic introduction, Springer-Verlag, BerlinHeidelberg

World Health Organisation, 2006: Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide, Global Update 2005. Geneva, 2006

World Health Organisation,2004: Environmental Epidemiology A Textbook on Study Methods and Public Health Applications. June 2004

World Health Organisation, 2003: Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide. Report on a WHO Working Group. Bonn, Germany, 13-15 January 2003

World Health Organisation,2001: Quantification of the health effects of exposure to air pollution. Report of a WHO working group. Bilthoven, Netherlands, 20-22 November 2000. Copenhagen, 2001